Letter to the Editor

Differentiating Between Selection and Mutation Bias

Adam Eyre-Walker

Centre for the Study of Evolution and Department of Biological Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom

Manuscript received December 21, 1996
Accepted for publication July 21, 1997

IT is often difficult to differentiate the effects of selection from mutation biases. For instance it is unclear whether the G+C rich parts of the mammalian genome, the so called G+C rich isochores, are maintained by mutation biases or selection (BERNARDI 1989; FILIPSKI 1990). BALLARD and KREITMAN (1994) and AKASHI (1995) have recently suggested a variation of the MacDonald-Kreitman (MACDONALD and KREITMAN 1991) test, which itself a variation of the HKA test (HUDSON *et al.* 1987), which can be used to test whether mutation biases are solely responsible for compositional biases in a sequence. The test involves comparing the pattern of substitution to the pattern of polymorphism. The bases at each site are first divided into two groups; these groups could be the G/C and A/T nucleotides, common and rare synonymous codons, or any other division of the four nucleotides at each site. Let us call the two groups of nucleotides preferred and unpreferred, for reasons that will become apparent. To conduct the test one tabulates the number of unpreferred to preferred, and preferred to unpreferred substitutions (s_p and s_u, respectively), and the number of preferred mutations segregating at sites that were unpreferred ancestrally, and the number of unpreferred mutations segregating at sites which were ancestrally preferred (m_p and m_u respectively); *i.e.*, m_p is the number of preferred mutations segregating in the population and s_p is the number of preferred mutations that have been fixed. Under neutrality AKASHI (1995), and BALLARD and KREITMAN (1994) suggest that the ratios s_p/s_u and m_p/m_u should be equal, a hypothesis that can be easily tested. This is clearly true if the system is stationary (*i.e.*, there is no overall change in the relative frequency of the two groups); when the system is stationary, the number of substitutions between the two groups must be equal (whether the sequence is under selection or not), and under neutrality the pattern of mutation must reflect the pattern of substitution; *i.e.*, both ratios are expected to be one. However the ratios are not expected to be equal if the system is not (or has not been) stationary, because as the base composition changes through time so the pattern of mutation changes. Furthermore, under neutrality the mutation pattern must have changed if the sequences are not stationary; if the change occurred after the time from which we are observing substitutions, then the pattern of substitution reflects two different mutation processes. For example imagine a segment of DNA that has recently gone through an increase in G+C content because of a change in the mutation pattern, but which is now stationary. There will be an excess of A:T → G:C substitutions but equal numbers of G:C and A:T mutations segregating in the population. Although AKASHI (1995) appreciated that there are problems if sequences are changing in composition, the problems were not explicitly stated. Here I illustrate the problem with a simple example and suggest approaches to identify and deal with datasets that are not stationary.

There are three possible outcomes of the test: the proportion of substitutions that are preferred can be equal to, greater than, or less than the proportion of mutations that are preferred. Let us define the preferred group to be the most frequent in the sequence that we are considering; these are the nucleotides preferred by either mutation or selection. For example, in a G+C-rich sequence the preferred group is the G and C nucleotides. In this case AKASHI (1995) has shown that under weak directional selection the proportion of substitutions that are preferred should be greater than the proportion of mutations that are preferred (*i.e.*, $s_p/s_u > m_p/m_u$) (when the system is stationary). This is likely to be true for other models of selection, such as stabilizing selection. I will also note here that gene conversion is indistinguishable from weak directional selection in these terms. In the following it is shown that s_p/s_u is often greater than m_p/m_u when the mutation pattern changes and that selection can be incorrectly inferred from the test.

Consider a series of sites at which selection is not acting. Let the mutation rate from preferred nucleotides to unpreferred be u and the mutation rate in the reverse direction be v (note we are implicitly assuming that u and v are constant; this is likely to be the case under many models). The change in the frequency of

Author e-mail: a.c.eyrewalker@sussex.ac.uk

the preferred nucleotides, f, can be modeled by the differential equation:

$$\frac{\partial f}{\partial T} = -fu + (1 - f)v \quad (1)$$

from which it follows that the frequency of the preferred nucleotides is

$$f(t) = w + (f_0 - w)e^{-t}, \quad (2)$$

time units after the population was at a frequency f_0, where $w = v/(u + v)$ and $t = (u + v)T$. The numbers of preferred and unpreferred mutations segregating are therefore

$$m_p(t) = \sum_{i=0}^{\infty} (1 - f(t - i))vk_i$$

$$m_u(t) = \sum_{i=0}^{\infty} f(t - i)uk_i \quad (3)$$

where k_i is the probability that a mutation that occurred i generations ago is still segregating in the population. In practical terms one will only have a sample of sequences; in this case k_i is the probability that one samples a mutation that occurred i generations in the past. For simplicity we will assume that the time between substitutions is long compared to the time mutations exist in a population; this is reasonable since $N(u + v)$ appears to be less than one for most organisms that have been studied except for RNA viruses [*i.e.*, the time for which each mutation segregates, generally less than N generations, is much less than the time between the fixation of different mutations, $1/(u + v)$ generations]. Under this assumption the composition does not change during the period in which mutations accumulate so we can simplify Equations 3 to

$$m_p(t) = (1 - f(t))vK$$

$$m_u(t) = f(t)uK \quad (4)$$

where

$$K = \sum_{i=0}^{\infty} k_i.$$

The proportion of segregating mutations which are preferred is then

$$z = \frac{(1 - f(t))w}{f(t) - 2wf(t) + w}, \quad (5)$$

where w is the strength of the mutation bias, f_0 is the initial frequency of preferred nucleotides, and t is the time in units of $(u + v)$. Let us also assume that fixation is instantaneous; again this is likely to be a good approximation when $N(u + v) \ll 1$ since the time to fixation is of the order of $4N$ generations. The numbers of preferred and unpreferred substitutions are

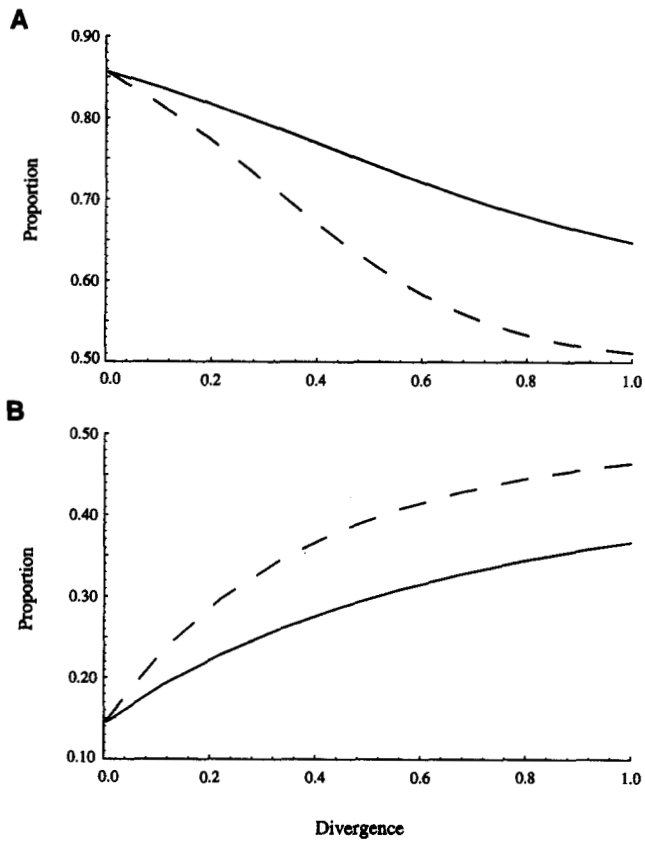

FIGURE 1.—The proportion of substitutions (——) and the proportion of segregating mutations (- - -) that are preferred nucleotides plotted against the overall proportion of sites that have changed. In part a the mutation bias has changed such that the preferred group is increasing (f_0 = 0.6, w = 0.9), whereas in part b the preferred group nucleotides are decreasing in frequency (f_0 = 0.9, w = 0.6).

$$S_p(t) = \int_0^t (1 - f(x))vdx$$

$$S_u(t) = \int_0^t f(x)udx, \quad (6)$$

which simplify to

$$s_p(t) = w\{(1 - w)t - (f_0 - w)(1 - e^{-t})\}$$

$$s_u(t) = (1 - w)\{wt + (f_0 - w)(1 - e^{-t})\}, \quad (7)$$

expressions in w, f_0 and t.

The proportion of substitutions that are preferred, $s_p/(s_p + s_u)$, and the proportion of mutations segregating in the population that are preferred, $m_p/(m_p + m_u)$, are plotted in Figure 1 against the proportion of sites that have changed. Two cases are illustrated. In Figure 1a the mutation pattern is such that the frequency of the preferred nucleotides is increasing. In this case s_p/s_u is greater than m_p/m_u; *i.e.*, under the Akashi/Ballard/Kreitman (ABK) test this sequence would appear to be subject to directional selection in favor of the preferred nucleotides. In Figure 1b the mutation pattern is such that the frequency of the pre-

ferred nucleotides is declining. In this case s_p/s_u is less than m_p/m_u; there is no simple selective explanation, although the result would not be consistent with naive neutral expectations. In both cases $s_p/su \neq m_p/m_u$ because the mutation pattern is changing as the composition of the sequence changes. The mutation pattern reflects the current mutation pattern while the substitution pattern reflects a changing process.

In both of these examples the change in the mutation pattern was assumed to have occurred at the point from which we are considering the pattern of substitution (the dynamics are similar if the change occurred before this point). This may not be the case, the pattern of mutation may have changed at some point more recently. Let us assume that the system was stationary prior to some time, αt, at which point the mutation bias changed instantly to a new value, w. Let the frequency of the preferred nucleotides prior to the change be f_0. If t time units have occurred since the change in the bias then

$$s_p^*(t) = f_0(1 - f_0)\alpha t + s_p(t)$$
$$s_u^*(t) = f_0(1 - f_0)\alpha t + s_u(t) \qquad (8)$$

m_p and m_u remain unchanged. The quantities $s_p^*/(s_p^* + s_u^*)$ and $m_p/(m_p + m_u)$ are plotted in Figure 2 for the case where the change in the mutation bias occurs two thirds of the way along the branch leading to the sequence being considered (*i.e.*, $\alpha = 2$). The dynamics are more complicated than when the change in the mutation bias occurred at or prior to the ancestral node. When the mutation bias changes along a branch such that the frequency of the preferred nucleotides increases, the proportion of substitutions that are preferred is less than the proportion of mutations that are preferred initially (*i.e.*, $s_p^*/s_u^* < m_p/m_u$) (Figure 2a); although it may take some considerable time, eventually this inequality is reversed (*i.e.*, $s_p^*/s_u^* > m_p/m_u$). In contrast if the mutation bias changes such that the frequency of the preferred nucleotides starts to decline, the proportion of substitutions that are preferred is less greater than the proportion of segregating mutations that are preferred initially (*i.e.*, $s_p^*/s_u^* > m_p/m_u$) (Figure 2b). In this case s_p^*/s_u^* and m_p/m_u are different because the substitution pattern reflects two mutation patterns; a stationary pattern and a new nonstationary pattern.

The pattern shown in Figure 2b is the situation most likely to lead to incorrect inferences about selection. In the other examples, the differences between s_p/s_u (or s_p^*/s_u^*) and m_p/m_u are relatively small and in the examples shown in Figures 1b and 2b the difference between the ratios is in a direction not expected under weak directional or stabilizing selection models unless there are strong mutation biases in operation as well. In contrast the differences between the ratios in Figure 2b is in the correct direction to be consistent with directional

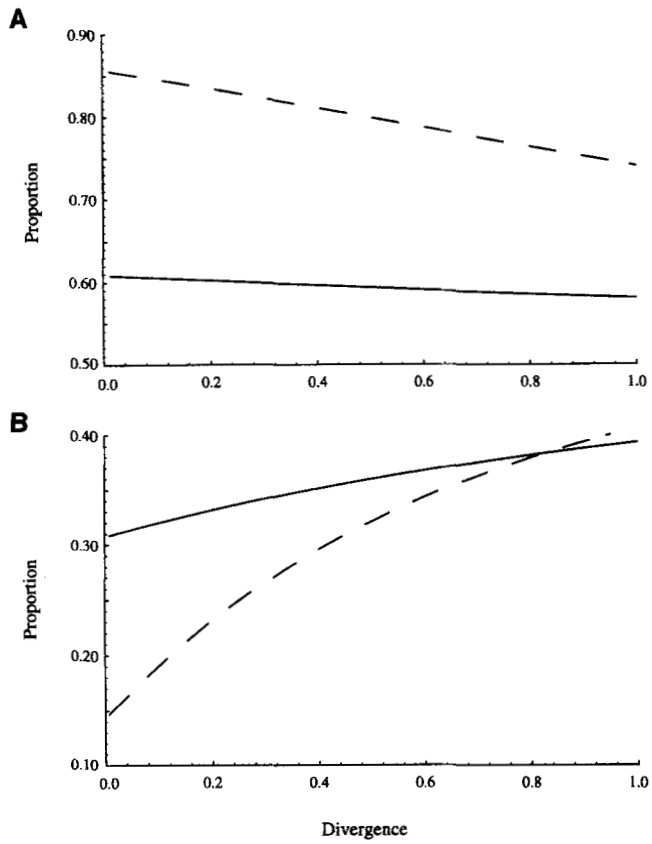

FIGURE 2.—The proportion of substitutions (——) and the proportion of segregating mutations (- - -) that are preferred nucleotides plotted against the overall proportion of sites that have changed, for the case where the mutation bias changes two thirds of the way along the lineage. In part a the mutation bias has changed such that the preferred group is increasing ($f_0 = 0.6$, $w = 0.9$), whereas in part b the preferred group nucleotides are decreasing in frequency ($f_0 = 0.9$, $w = 0.6$).

and stabilizing selection and can be made as extreme as required; the more recently the mutation bias has changed, the greater the difference between the substitution pattern (s_p^*/s_u^*) and the mutation pattern (m_p/m_u).

In general one should therefore be cautious about applying the ABK test unless the system appears to be stationary. It is important to appreciate that since the test is designed to differentiate between selection and mutation, a change in the mutation bias is an equally parsimonious explanation of why the sequences are not stationary as a change in the strength or direction of selection.

There are three solutions to the "stationary" problem; the first is to test whether the system is stationary; the second is to use an alternative test based on the frequency distribution of segregating mutations, and the third is to determine what change in the mutation pattern would be required to explain the data. Whether the sequences are stationary can be tested using the substitution data. Ideally we would like to be able to test

whether the number of preferred substitutions along a lineage is equal to the number of unpreferred substitutions since $s_p = s_u$ when the sequences are stationary. Unfortunately this is not straightforward since the reconstruction of ancestral states using parsimony is problematic in sequences of biased composition (COLLINS *et al.* 1994; PERNA and KOCHER 1995). Problems arise because the rate of preferred to unpreferred substitutions (per site) is lower than the rate of unpreferred to preferred substitutions when a sequence is stationary; this means that sites that were ancestrally unpreferred change to sites that are preferred in two of three taxa much more rapidly than the reverse process; *i.e.*, the rate of UUU → PUP is greater than PPP → UPU. An alternative test is to use a single outgroup and compare the number of sites that are fixed for a preferred nucleotide in the ingroup and an unpreferred nucleotide in outgroup (d_{pu}) and vice versa (d_{up}). If the sequences are stationary these are expected to be equal even if one of the lineages evolves faster than the other and there are multiple substitutions (EYRE-WALKER 1994).

The ABK test is therefore best split into two components: a test of whether the system is stationary using the substitution data (*i.e.*, $d_{up} = d_{pu}$), and a test for selection using the polymorphism data (*i.e.*, $m_p = m_u$). This will a more powerful test if the system is stationary since there is only one source of sampling error in the test of selection; however the test of whether the sequences are stationary may not be very powerful so caution should always be exercised when interpreting the results.

An alternative is to test the frequency distributions of preferred and unpreferred mutations against one another (SAWYER *et al.* 1987; AKASHI and SCHAEFFER 1997; R. KLIMAN, personal communication). Under neutrality, if the sequences are stationary, then the frequency distributions of preferred and unpreferred mutations should be the same. Furthermore this will be true if the sequences are not stationary so long as the mutation pattern has not changed within the last $\sim 4N$ generations. In contrast the ABK test is sensitive to changes in the mutation pattern within $\sim 1/(u + v)$ generations. Since $4N(u + v) \ll 1$ in most organisms, the ABK test is more sensitive to changes in the mutation pattern than the frequency distribution test.

Whether or not there is evidence that the sequences are stationary, it may be useful to estimate the change in the mutation pattern required to explain the data; if the change is extreme one may doubt a mutational explanation, especially if a simple selection model is consistent with the data. The change in the mutation pattern can be estimated in the following way; the original mutation bias is estimated from the present frequency of preferred nucleotides: *i.e.*, $w^* = f$. This is a conservative estimate. The new mutation bias is then

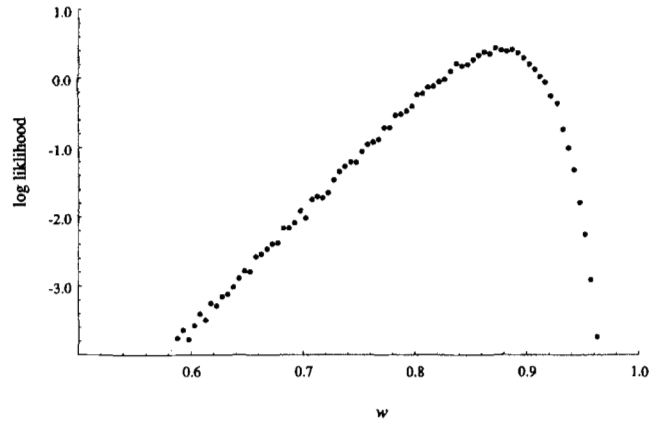

FIGURE 3.—Likelihood surface for the new mutation bias, w, for the *Drosophila cytochrome b* data of BALLARD and KREITMAN (1994). 100,000 sets of parameter values (y_i) were generated and the likelihood was summed over intervals of 0.005.

estimated from the level of bias and the pattern of polymorphism by rearranging Equation 5:

$$w = \frac{fz}{1 - z - f + 2fz}. \tag{9}$$

An unbiased estimate of w and its confidence intervals can be obtained by bootstrapping, unless there are no preferred mutations segregating (*i.e.*, $z = 0$), or by maximum likelihood. There are four types of site in the analysis, preferred and unpreferred sites without (unpreferred and preferred respectively) mutations segregating, and preferred sites with unpreferred mutations segregating, and unpreferred sites with preferred mutations segregating. These are multinomially distributed. Let the observed numbers of the four types of site be x_1, $x_2 \cdots$ etc.; *e.g.*, $x_3 = m_u$. The likelihood of observing x_i given that the true proportions of the four types of site are y_i is

$$L = \frac{(\Sigma x_i)!}{\Pi x_i!} \Pi y_i^{x_i}. \tag{10}$$

Thus by randomly generating y_i between 0 and 1 such that $\Sigma y_i = 1$, and substituting the values y_i into Equation 9, it is possible to construct the likelihood surface for w. In practice I have found that generating values of y_i within four standard errors of their observed values gives a good approximation to the likelihood surface. 95% confidence intervals can then be inferred from a decline in the log likelihood of two units.

BALLARD and KREITMAN (1994) and AKASHI (1995) used parsimony to determine the direction of substitutions and mutations segregating in the population. This method is efficient when divergences between the sequences are low but becomes biased in favor of preferred to unpreferred substitutions and mutations when divergences are moderate or large (COLLINS *et al.* 1994; PERNA and KOCHER 1995). However the direction of

mutation can be determined from the frequencies of the segregating alleles; sites at which the preferred allele is in a minority are inferred to be a preferred mutation segregating at an unpreferred site, and vice versa. This will be unbiased under the null hypothesis if the change in mutation pattern occurred more than $\sim 4N$ generations ago since the frequency distributions of preferred and unpreferred mutations are expected to be the same if the composition of the sequence is solely determined by mutation biases. If the change in the mutation pattern occurred within $\sim 4N$ generations of the present the estimate will be conservative.

To illustrate the principles laid out here, let us reconsider whether synonymous codon bias has been maintained by mutation in the mitochondrial cytochrome *b* gene of Drosophila. BALLARD and KREITMAN (1994) presented sequence data from 17 lines of *Drosophila melanogaster*, 18 lines of *Drosophila simulans* and 14 lines of *Drosophila yakuba*; for the purposes of this analysis I have followed BALLARD and KREITMAN and ignored the single *sI D. simulans* line since this appears to be quite distinct from other *simulans* lines. The cytochrome *b* gene is very AT rich at synonymous sites. Of the 15 G:C ↔ A:T synonymous site polymorphisms segregating in the three species, 11 have GC segregating at a frequency of < 0.5, so we infer that 11 of the polymorphisms arose via A:T → G:C mutations, 4 by G:C → A:T mutations; the difference is nearly significant ($P < 0.10$ in a one-tail binomial test). There is no evidence that the sequences are changing in composition—there are 13 GC_{mel}:AT_{sim} to 21 AT_{mel}:GC_{sim}, 20 GC_{mel}:AT_{yak} to 21 AT_{mel}:GC_{yak}, and 25 GC_{sim}:AT_{yak} to 22 AT_{sim}:GC_{yak} fixed differences—so the results are consistent with selection acting in favor of AT. However if we calculate the change in the mutation pattern required to explain the data, it turns out to be small. The average AT content at synonymous sites in the cytochrome *b* gene is 94%, so the original mutation bias prior to the change, w^*, must have been at least 0.940 with 95% CIs of 0.916 to 0.964. The maximum likelihood estimate for the new mutation pattern is ~0.88 (Figure 3); this is the G+C content that we would expect when the sequences are stationary. The 95% CIs are ~0.72 to ~0.95. There is therefore little evidence of selection upon synonymous codon use in the Drosophila cytochrome *b* gene.

The ABK procedure tests whether a compositional bias is due to mutation alone; it is therefore likely to be useful in detecting selection that causes compositional bias and selection in sequences subject to a mutation bias. Since compositional biases are widespread, the test may have broad application; the compositional biases must be due to selection or mutation; in either case the ABK test may be able to detect the action of selection. Like the test of MACDONALD and KREITMAN (1991) the ABK test makes few assumptions; because preferred and unpreferred sites are interspersed along the sequence it does not assume that the sites are independent or that population sizes have been constant. Furthermore, departures from neutrality can be attributed directly to selection acting upon the sites under consideration, rather than selection at linked loci.

I am very grateful to RICH KLIMAN, HIROSHI AKASHI, BRANDON GAUT, JOHN WAKEFIELD, JODY HEY, ANDY CLARK and two anonymous referees for helpful discussion and comments on this manuscript.

LITERATURE CITED

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

AKASHI, A., and S. W. SCHAEFFER, 1997 Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics **146:** 295–307.

BALLARD, J. W. O., and M. KREITMAN, 1994 Unraveling selection in the mitochondrial genome of Drosophila. Genetics **138:** 757–772.

BERNARDI, G., 1989 The isochore organization of the human genome. Annu. Rev. Genet. **23:** 637–661.

COLLINS, T. M., P. H. WIMBERGER and G. J. P. NAYLOR, 1994 Compositional bias, character state bias and character state reconstruction using parsimony. Syst. Biol. **43:** 482–496.

EYRE-WALKER, A., 1994 DNA mismatch repair and synonymous codon evolution in mammals. Mol. Biol. Evol. **11:** 88–98.

FILIPSKI, J., 1990 Evolution of DNA sequence. Contributions of mutational bias and selection to the origins of chromosomal compartments, pp. 1–54 in *Advances in Mutagenesis*, Vol. 2, edited by G. OBE. Springer-Verlag, New York.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

MACDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature **351:** 652–654.

PERNA, N. T., and T. D. KOCHER, 1995 Unequal base frequencies and the estimation of substitution rates. Mol. Biol. Evol. **12:** 359–361.

SAWYER, S., D. E. DYKHUIZEN and D. L. HARTL, 1987 Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc. Natl. Acad. Sci. USA **84:** 6225–6228.

Communicating editor: A. G. CLARK